Phylogenomic conflict analyses in the apple genus *Malus* s.l. reveal widespread hybridization and allopolyploidy driving diversification, with insights into the complex biogeographic history in the Northern Hemisphere

Bin-Bin Liu[1,2], Chen Ren[3,4], Myounghai Kwak[5], Richard G.J. Hodel[2], Chao Xu[1], Jian He[6], Wen-Bin Zhou[7], Chien-Hsun Huang[8], Hong Ma[9], Guan-Ze Qian[10], De-Yuan Hong[1*], Jun Wen[2*]

SUPPORTING INFORMATION

Supplementary Methods

Single-copy nuclear marker development

The SCN marker development followed the pipeline in Liu et al. (2021). Briefly, the coding regions of *Malus domestica* (GenBank assembly accession: GCA_000148765.2) were first input into MarkerMiner v.1.0 (Chamala et al. 2015) to identify the putative single-copy genes. The resulting genes were then filtered by successively BLASTing (Altschul et al. 1990, 1997; Camacho et al., 2009) against six available genomes [*Malus baccata* (accession: GCA_006547085.1), *M. domestica*, *Pyrus betulifolia* Bunge (accession: GCA_007844245.1), *P. bretschneideri* Rehder (accession: GCA_000315295.1), *P. ussuriensis* Maxim. × *P. communis* L. (accession: GCA_008932095.1), and *P. pyrifolia* (Burm.f.) Nakai (accession: GCA_016587475.1)] in Geneious Prime (Kearse et al., 2012), with the parameters settings in the Megablast program (Morgulis et al., 2008) as a maximum of 60 hits, a maximum E-value of 1×10^{-10}, a linear gap cost, a word size of 28, and scores of 1 for match and -2 for mismatch in alignments. We first excluded genes with mean coverage > 1.1 for alignments, which generally indicate potential paralogy of the genes and/or the presence of highly repeated elements in the sequences. The remaining alignments were further visually examined to exclude those genes receiving multiple hits with long overlapping but different sequences during the BLAST. It should be noted that the alignments with mean coverage between 1.0 and 1.1 were generally caused by the presence of tiny pieces of flanking intron sequences in the alignments. These fragments were still accepted as a SCN gene here. After filtering, the remaining genes were used as references in the following gene assembly. The baits *in silico* could be available from the Dryad Digital Repository (Data 1): https://doi.org/10.5061/dryad.2jm63xsq5.

Data processing and the assembly of single-copy nuclear genes

Read processing and assembly followed the pipeline in Liu et al. (2021). Generally, we

used Trimmomatic v. 0.39 (Bolger et al. 2014) for quality trimming and adapter clipping with the parameters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). Then, the results were quality-checked with FastQC v. 0.11.9 (Andrews 2018). The number of clean reads after trimming was also calculated here for comparison (Supplementary Table S1). We used the HybPiper pipeline v. 1.3.1 (Johnson et al., 2016) for targeting SCN genes with default settings; BWA v. 0.7.1 (Li & Durbin 2009) to align and distribute reads to target genes; SPAdes v. 3.15.0 (Bankevich et al., 2012) with a coverage cutoff value of 8 to assemble reads to contigs; and Exonerate v. 2.2.0 (Slater & Birney 2005) to align assembled contigs to target sequences and determine exon-intron boundaries. Python and R scripts included in the HybPiper pipeline (Johnson et al., 2016) were used to retrieve the recovered gene sequences, summarize and visualize the recovery efficiency.

Plastome assembly, annotation, phylogenetic analysis, and discordance analysis

A two-step strategy was used for obtaining high-quality chloroplast genomes. NOVOPlasty v. 4.3.1 (Dierckxsen et al., 2016) was applied first to assemble the plastomes with high-quality raw data. Then we used the successive assembly approach (Zhang et al. 2015), combining the reference-based and the *de novo* assembly methods to assemble the remaining low-quality samples. With the *de novo* assembly and a seed-and-extend algorithm, NOVOPlasty was the least laborious approach and resulted in accurate plastomes; however, this program needs sufficient high-quality raw reads without gaps to cover the whole plastome. The whole plastomes assembled from NOVOPlasty then could be used as references for assembling the remaining samples. The successive method provided an excellent approach to obtaining relatively accurate and nearly complete plastomes with or without gaps from lower-coverage raw data. Due to the sensitivity of Bowtie2 v. 2.4.2 (Langmead et al., 2012) to the reference, this successive method needs a closely related reference sequence with increased time and RAM requirements. Several recent studies have described the procedure in detail (Liu et al.

2019, 2020a, 2020b, 2021, Wang et al. 2020). All assembled plastomes have been submitted to GenBank with the accession numbers listed in Supplementary Table S1.

The assembled plastid genomes from the low-coverage and high-coverage datasets were annotated using PGA (Qu et al., 2019) with a closely related plastome (MN062004: *Malus ioensis*) downloaded from GenBank as the reference, and the results of automated annotation were checked manually. The coding sequences of plastomes were translated into proteins to manually check the start and stop codons in Geneious Prime (Kearse et al. 2012). The custom annotations in the GenBank format were converted into the FASTA format, and five-column feature tables file required by NCBI submission using GB2sequin (Lehwark & Greiner 2019).

Given the considerable variation among plastid introns at the level of Maleae, we extracted 80 coding genes (CDSs) using Geneious Prime (Kearse et al. 2012), and these CDSs were aligned by MAFFT v. 7.475 (Nakamura et al. 2018) with default parameters. This dataset with 77 samples and 80 plastid coding genes is available from the Dryad Digital Repository (Data 5): https://doi.org/10.5061/dryad.2jm63xsq5. The best-fit partitioning schemes and/or nucleotide substitution models for each dataset were estimated using PartitionFinder2 (Stamatakis, 2006; Lanfear et al., 2016), under the corrected Akaike information criterion (AICc) and linked branch lengths, as well as with rcluster (Lanfear et al. 2014) algorithm options. The partitioning schemes and evolutionary model for each subset were used for the downstream phylogenetic analysis. Like the nuclear analysis, we estimated the ML tree by IQ-TREE2 v. 2.1.3 (Minh et al. 2020) with 1000 SH-aLRT and the ultrafast bootstrap replicates and RAxML 8.2.12 (Stamatakis 2014) with GTRGAMMA model for each partition and clade support assessed with 200 rapid BS replicates. We also used ASTRAL-III (Zhang et al. 2018) for estimating a coalescent-based species tree. These three trees are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.2jm63xsq5.

A general procedure for assessing phylogenomic discordance and biogeographic analysis using deep genome skimming data

1. Two datasets (SCN genes and plastid coding genes) from deep genome skimming were generated for phylogenomic discordance and historical biogeographic analyses.
2. Given the sample's uneven and limited sequencing coverage, the outlier sites, missing data, and short sequences in the assembled SCN genes need to be trimmed using several programs, such as trimAL, Spruceup, and TreeShrink.
3. Generate nested datasets based on the number of genes and samples and then perform the phylogenetic analysis using concatenated supermatrix (RAxML and/or IQ-TREE2) and coalescent-based methods (ASTRAL-III, SVDquartets, MP-EST, Quartet MaxCut, etc.). The potential topology conflicts could provide some insights into the introgression between species, and the alternative relationship tested in *phyckle* could sort out the gene trees supporting each bipartition.
4. Detect and visualize gene tree discordance using *phyparts* and/or Quartet Sampling.
5. If no conflict exists among topologies and gene trees, any topology could be used for the following analysis. If conflicts are shown, evaluate the possible causes of these conflicts.
6. Observed phylogenomic discordance may be due to incomplete lineage sorting, allopolyploidy, and hybridization. The coalescence simulation could measure the goodness-of-fit of the coalescent model with ILS explaining the gene tree discordance adequately. When ILS could not explain the discordance sufficiently, phylogenetic network analyses are appropriate for testing hybridization, paying attention to the number of terminals (not more than 30) for evaluation. If the number is more than 30, subsampling the species for phylogenetic network analysis is necessary. Successive network analyses using different datasets at various taxonomic levels may be adopted as an alternative strategy on the limitation of the number of species. With chromosome

count data included, the role of allopolyploidy in explaining phylogenomic conflict may be evaluated.

7. Historical biogeographic analysis is performed, including divergence time and ancestral area estimation. It should be noted that the fossil species represent the extinct species in the evolutionary history of the focal lineages, hence historical biogeographic analysis involved in fossil and living species will promote a more accurate divergence time estimation and ancestral area reconstruction.
8. Possible cytonuclear discordance may also exist in some angiosperm lineages, and different topologies could significantly affect the ancestral area estimation. The divergence time estimation may also show conflicts. Generally, the overall age estimates in the nuclear data appear to be older than those estimated from the plastid coding genes, and this has been confirmed in our case study and some recent studies (e.g., Dong et al. 2021; Wang et al. 2021). The cytonuclear discordance could be explained by ILS and/or chloroplast capture.
9. Further test when and where the hybridization, allopolyploidy, and chloroplast capture events may have occurred.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic Local Alignment Search Tool. J. Mol. Biol. **215**: 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**: 3389-3402.

Andrews, S. (2018). FastQC: A quality control tool for high throughput sequence data. [accessed August 1, 2021].

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., and Prjibelski, A.D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**: 455-477.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114-2120.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics **10**: 421.

Chamala, S., García, N., Godden, G.T., Krishnakumar, V., Jordon-Thaden, I.E., De Smet, R., Barbazuk, W.B., Soltis, D.E., and Soltis, P.S. (2015). MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. Appl. Plant Sci. **3**: 1400115.

Dierckxsens, N., Mardulyn, P., and Smits, G. (2016). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. **45**: e18-e18.

Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J.C., and Wickett, N.J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. **4**: 1600016.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., and Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics **28**: 1647-1649.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., and Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. **14**: 82.

Lanfear, R., Frandsen, P.B., Wright, A.M., Senfeld, T., and Calcott, B. (2016). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. **34**: 772-773.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**: 357.

Lehwark, P., and Greiner, S. (2019). GB2sequin-A file converter preparing custom GenBank files for database submission. Genomics **111**: 759-761.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**: 1754-1760.

Liu, B.B., Hong, D.Y., Zhou, S.L., Xu, C., Dong, W.P., Johnson, G., and Wen, J. (2019). Phylogenomic analyses of the *Photinia* complex support the recognition of a new genus *Phippsiomeles* and the resurrection of a redefined *Stranvaesia* in Maleae (Rosaceae). J. Syst. Evol. **57**: 678-694.

Liu, B.B., Campbell, C.S., Hong, D.Y., and Wen, J. (2020a). Phylogenetic relationships and chloroplast capture in the *Amelanchier-Malacomeles-Peraphyllum* clade (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data using genome skimming. Mol. Phylogenet. Evol. **147**: 106784.

Liu, B.B., Liu, G.N., Hong, D.Y., and Wen, J. (2020b). *Eriobotrya* belongs to *Rhaphiolepis* (Maleae, Rosaceae): evidence from chloroplast genome and nuclear ribosomal DNA data. Front. Plant Sci. **10**: 1731.

Liu, B.B., Ma, Z.Y., Ren, C., Hodel, R.G.J., Sun, M., Liu, X.Q., Liu, G.N., Hong, D.Y., Zimmer, E.A., and Wen, J. (2021). Capturing single‐copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: A case study in Vitaceae. J. Syst. Evol. **59**: 1124-1138.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. **37**: 1530-1534.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R., and Schäffer, A.A. (2008). Database indexing for production MegaBLAST searches. Bioinformatics **24**: 1757-1764.

Nakamura, T., Yamada, K.D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics **34**: 2490-2492.

Qu, X.J., Moore, M.J., Li, D.Z., and Yi, T.S. (2019). PGA: a software package for rapid, accurate, and flexible batch annotation of plastomes. Plant Methods **15**: 50.

Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC bioinformatics **6**: 31.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**: 2688-2690.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**: 1312-1313.

Wang, H.X., Morales-Briones, D.F., Moore, M.J., Wen, J., and Wang, H.F. (2021). A phylogenomic perspective on gene tree conflict and character evolution in Caprifoliaceae using target enrichment data, with Zabelioideae recognized as a new subfamily. J. Syst. Evol. **59**: 897-914.

Wang, Y.B., Liu, B.B., Nie, Z.L., Chen, H.F., Chen, F.J., Figlar, R.B., and Wen, J. (2020). Major clades and a revised classification of *Magnolia* and Magnoliaceae based on whole plastid genome sequences via genome skimming. J. Syst. Evol. **58**: 673-695.

Zhang, N., Wen, J., and Zimmer, E.A. (2015). Congruent deep relationships in the grape family (Vitaceae) based on sequences of chloroplast genomes and mitochondrial genes via genome skimming. Plos One **10**: e0144701.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform. **19**: 153.

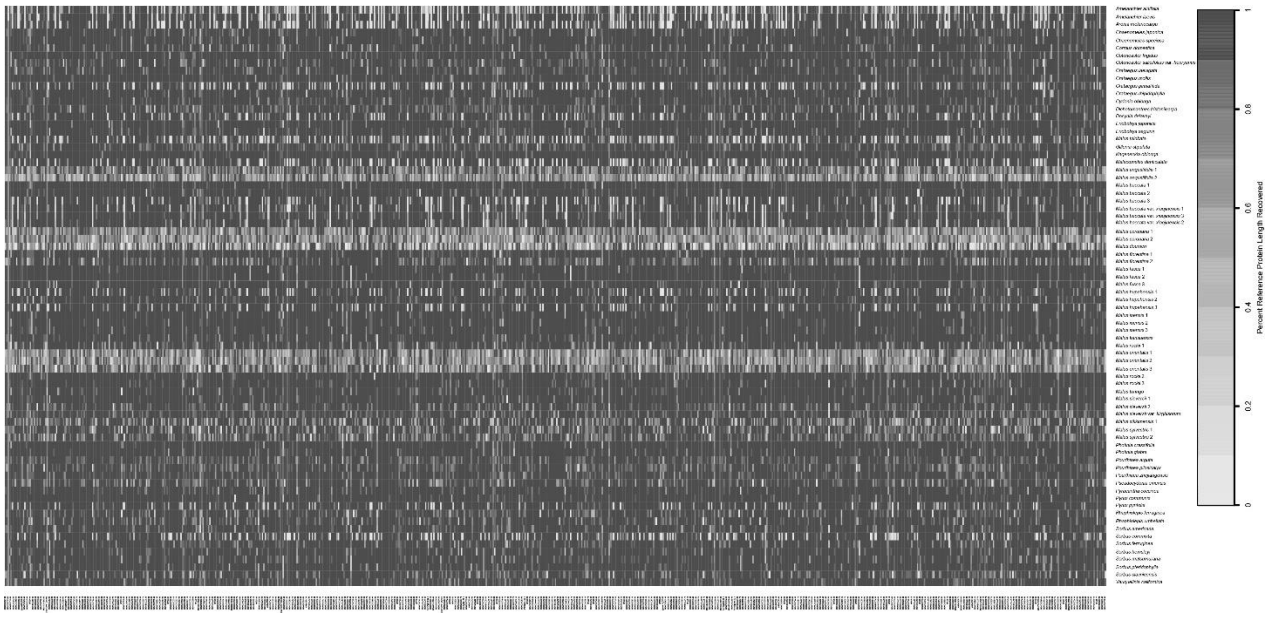

Figure S1. Heat map showing recovery efficiency for 797 genes recovered by HybPiper
Each column is a gene, and each row is one sample. The shade of gray in the cell is determined by the length of sequence recovered by the pipeline, divided by the length of the reference gene (maximum of 1.0).

218

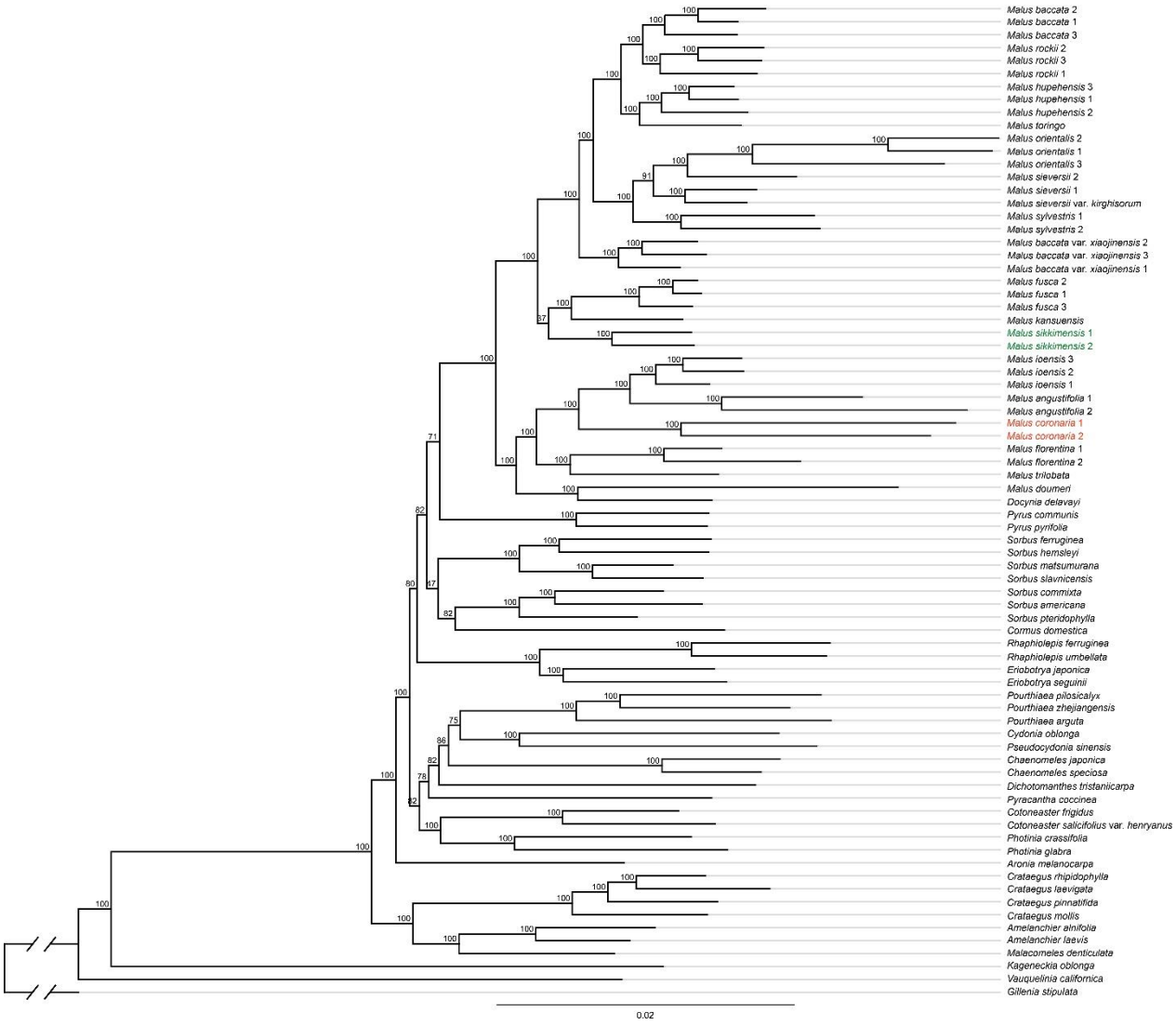

Figure S2. Maximum likelihood phylogeny of Malus s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 50%-sample dataset

Numbers above the branches indicate the bootstrap support.

219

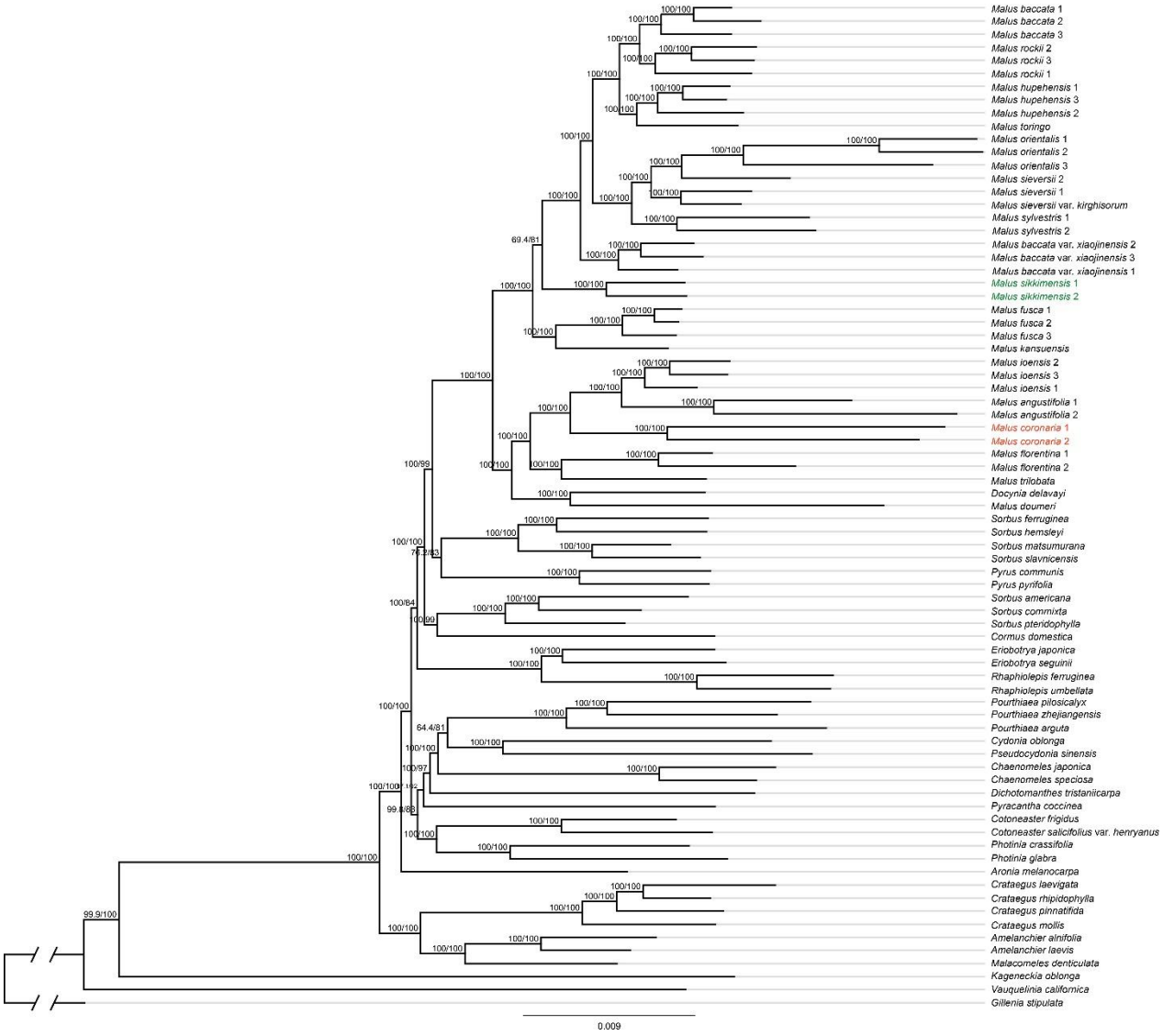

Figure S3. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the concatenated 50%-sample dataset

Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap support.

220

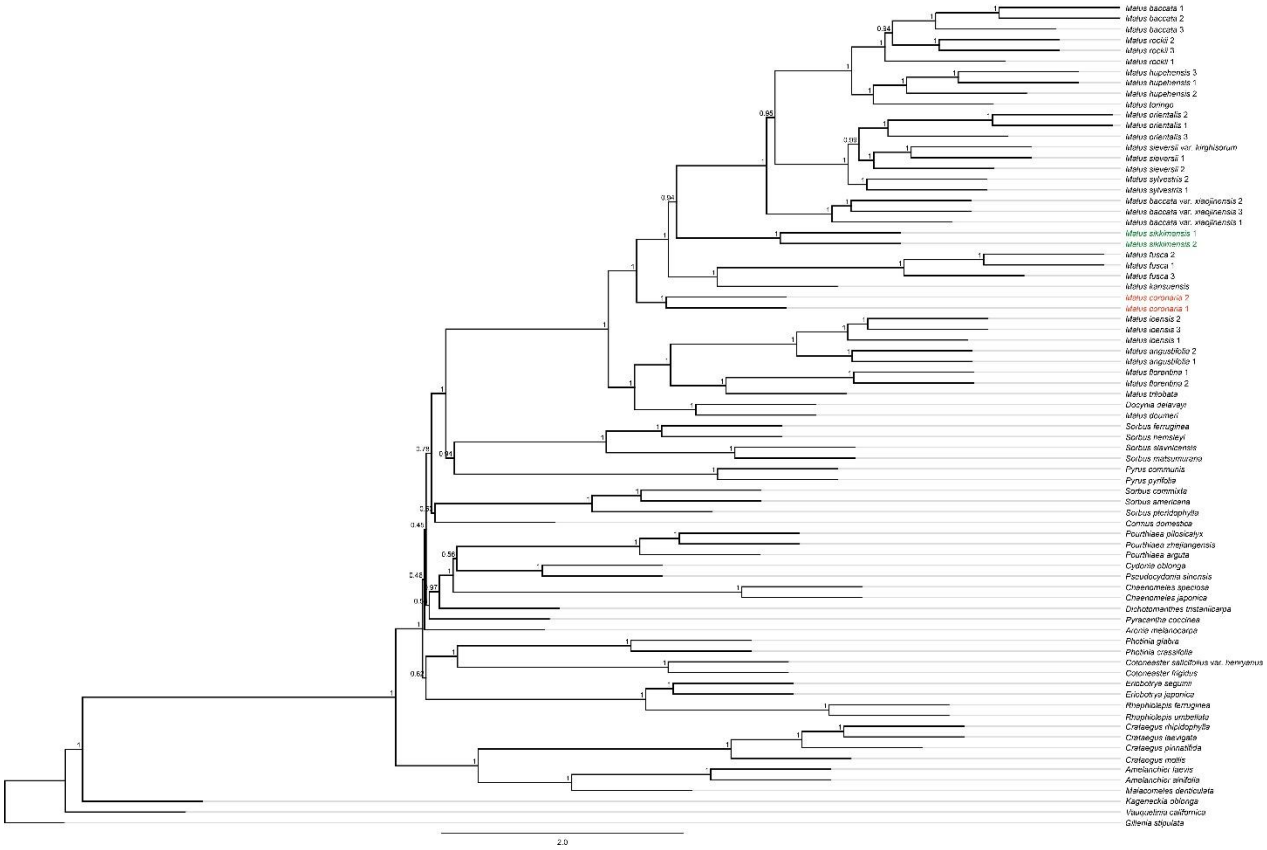

Figure S4. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the concatenated 50%-sample dataset

Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

221

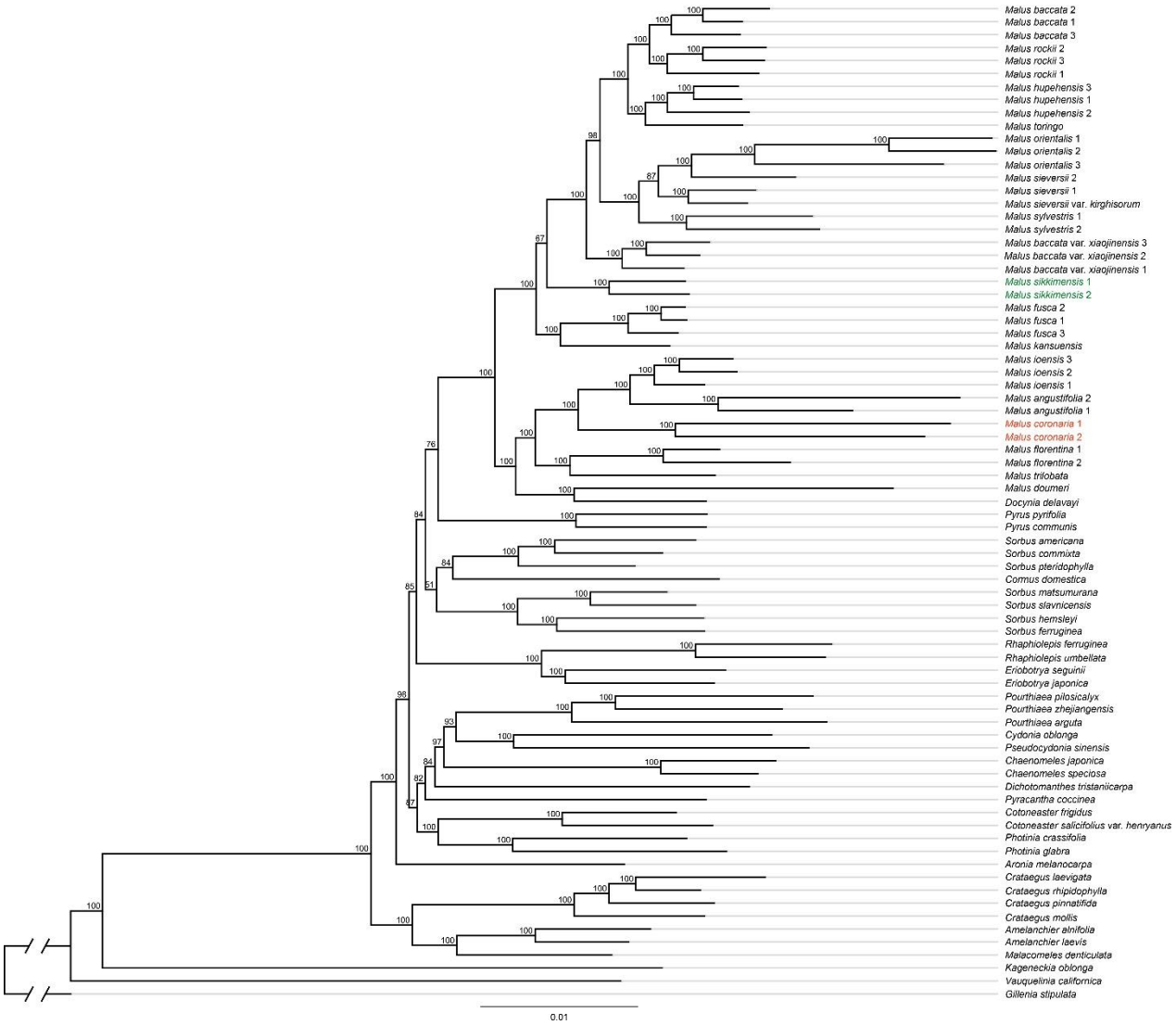

Figure S5. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80%-sample dataset

Numbers above the branches indicate the bootstrap support (BS).

222

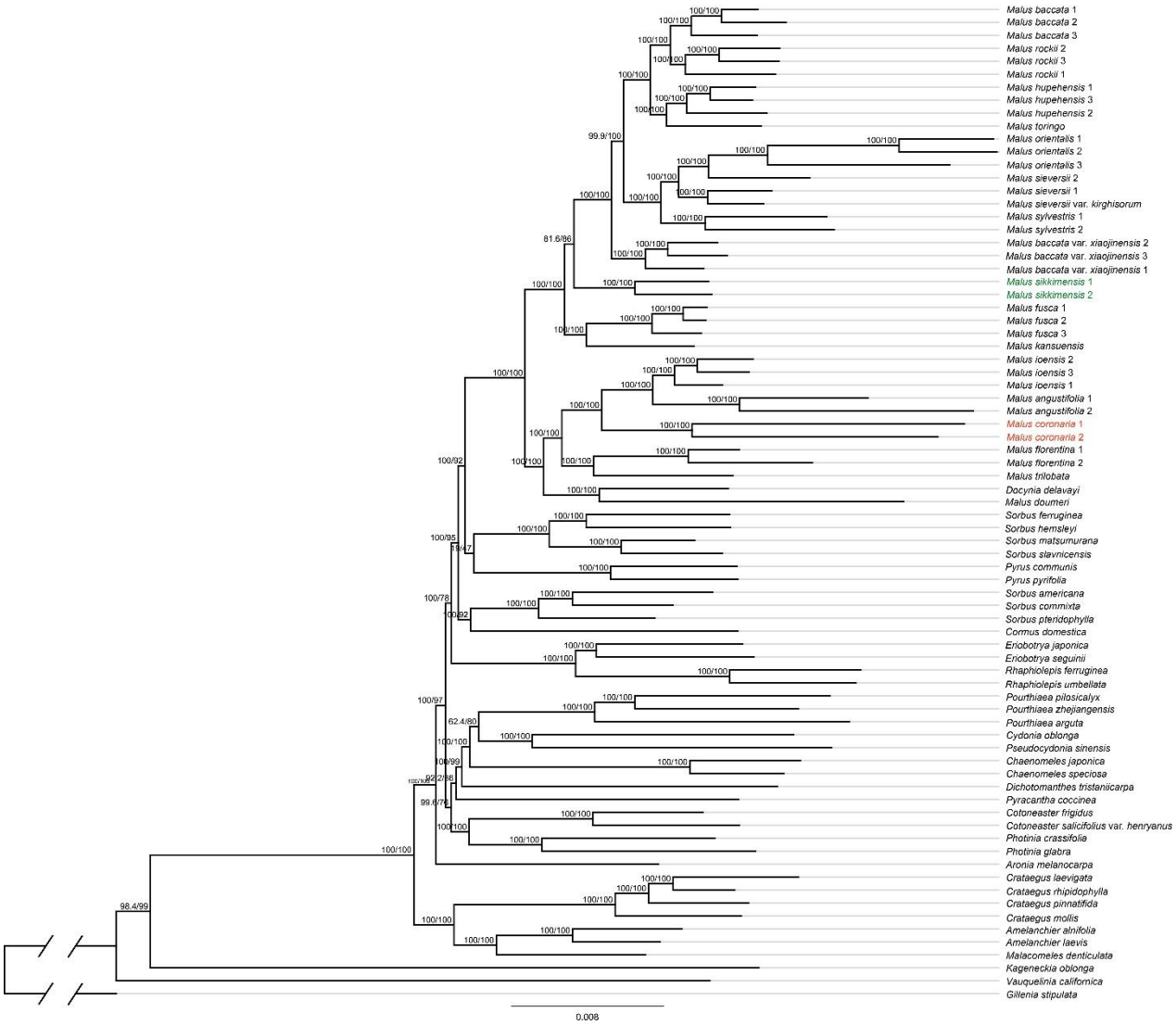

Figure S6. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the concatenated 80%-sample dataset

Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap (UFBoot) support.

223

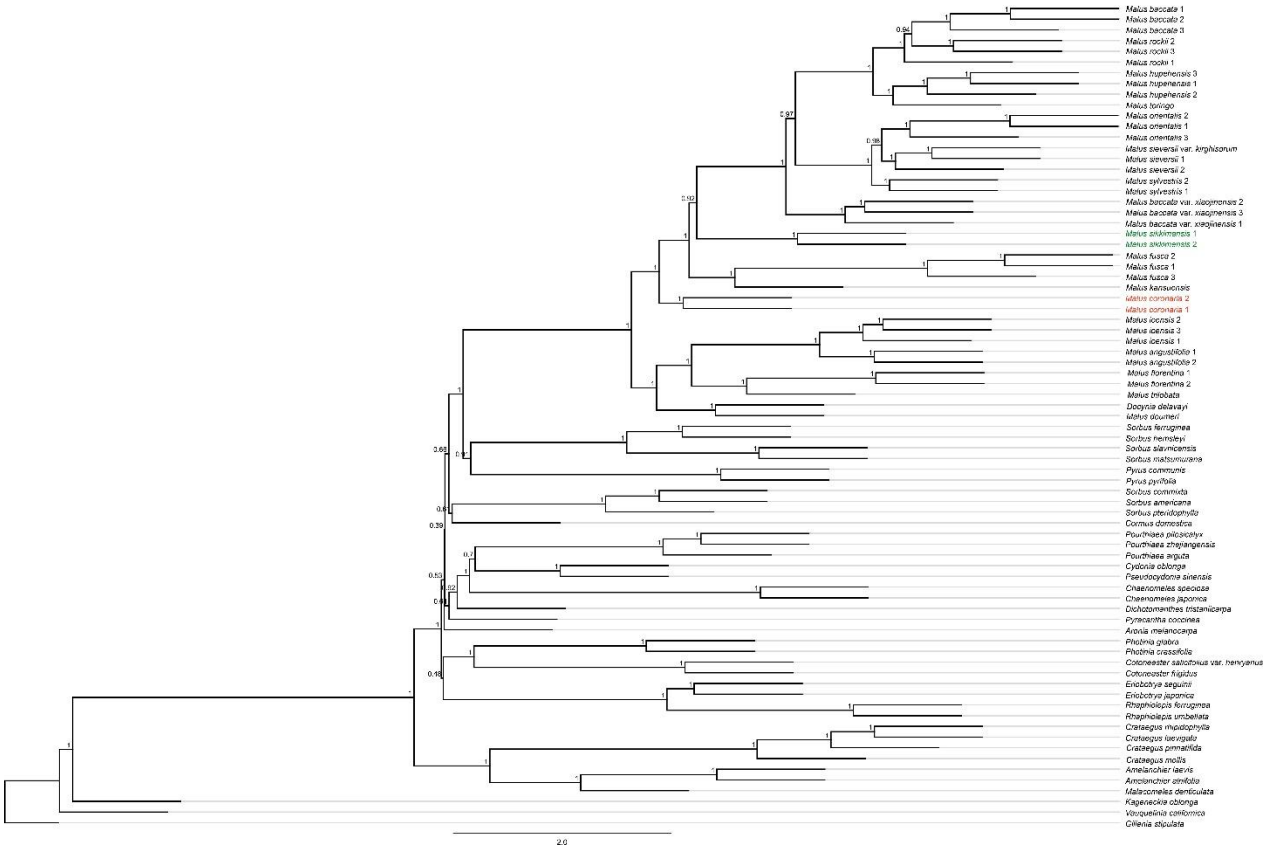

Figure S7. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the concatenated 80%-sample dataset

Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

224

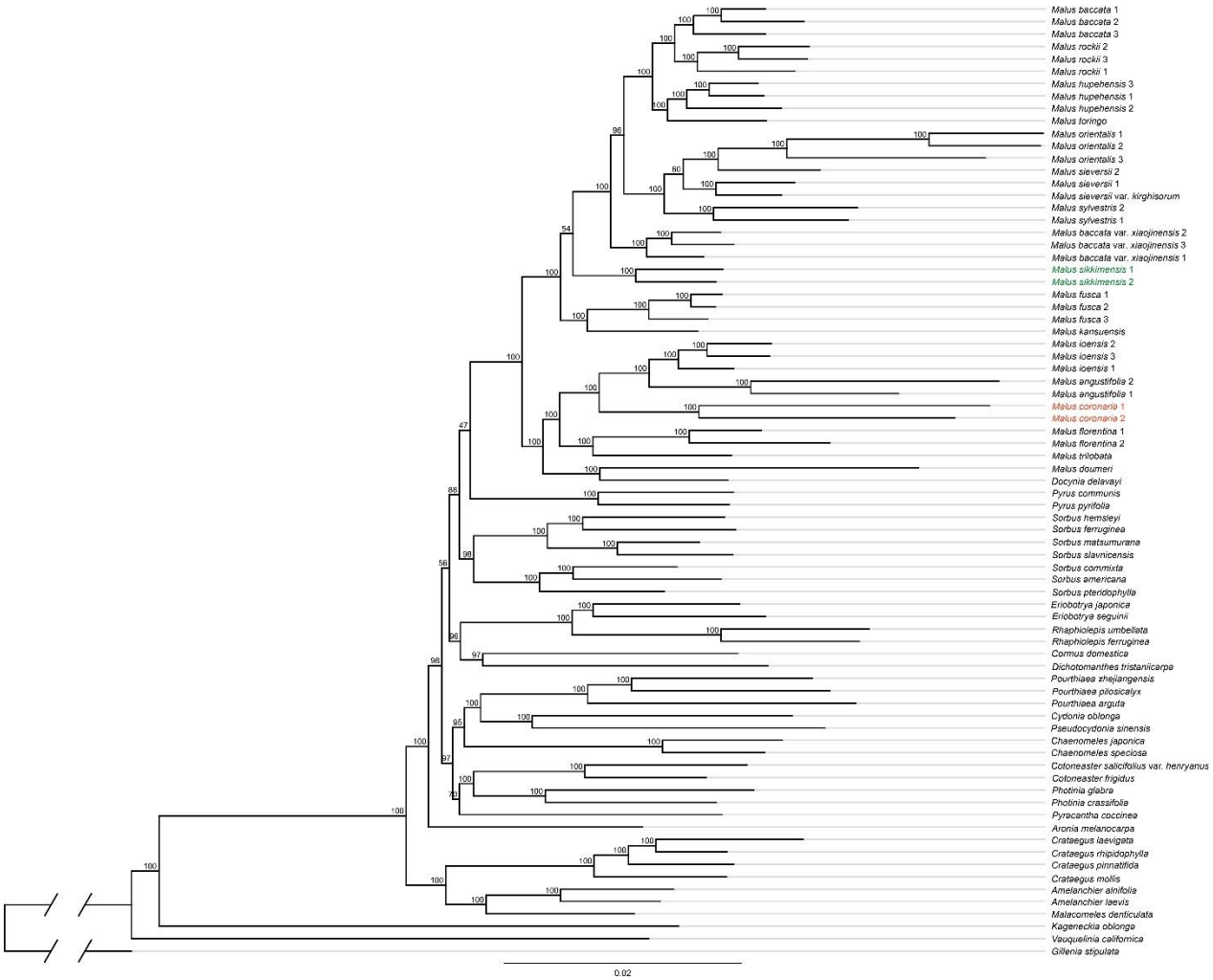

Figure S8. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the all-sample dataset

Numbers above the branches indicate the bootstrap support (BS).

225

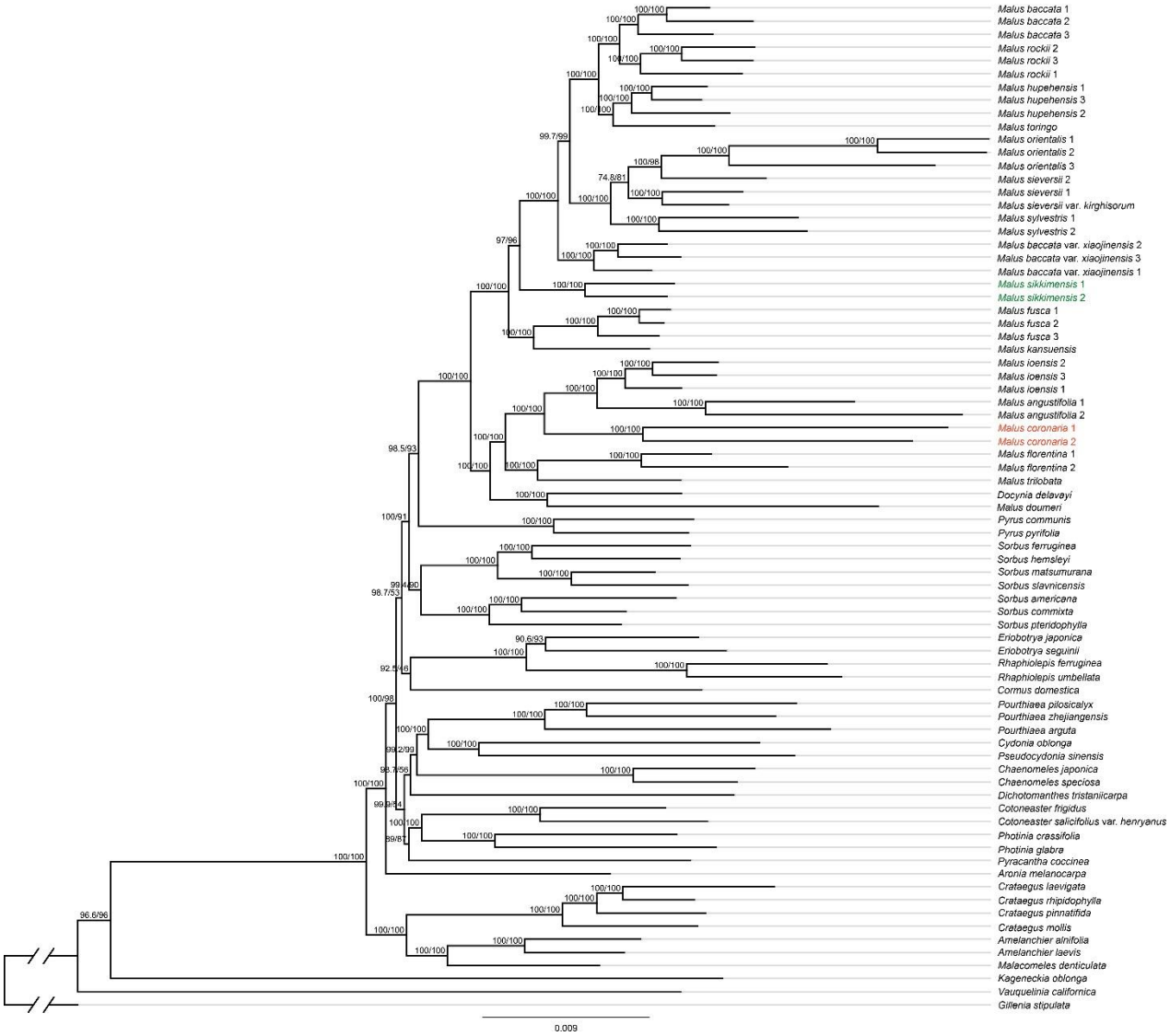

Figure S9. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the all-sample dataset

Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap (UFBoot) support.

226

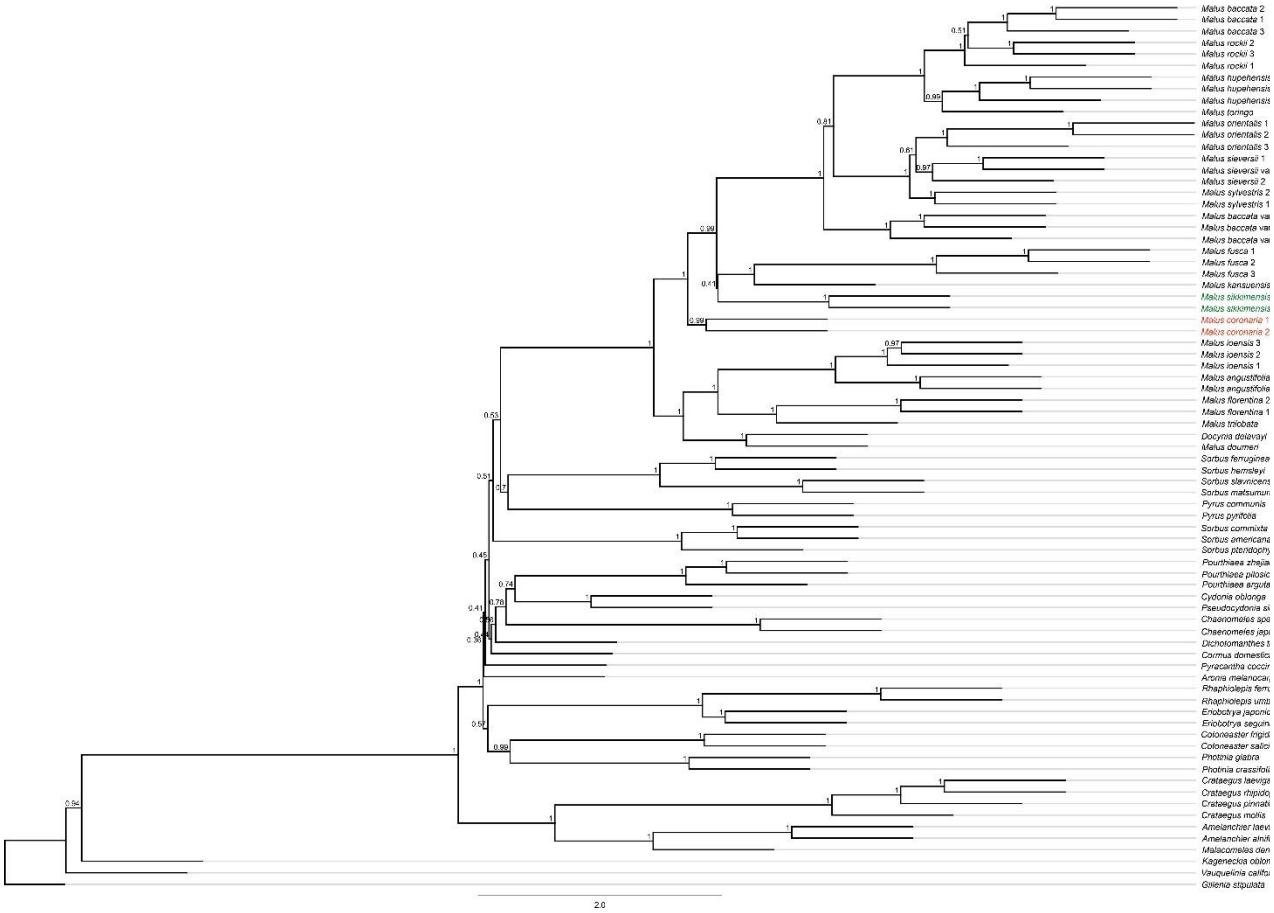

Figure S10. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the all-sample dataset

Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

227

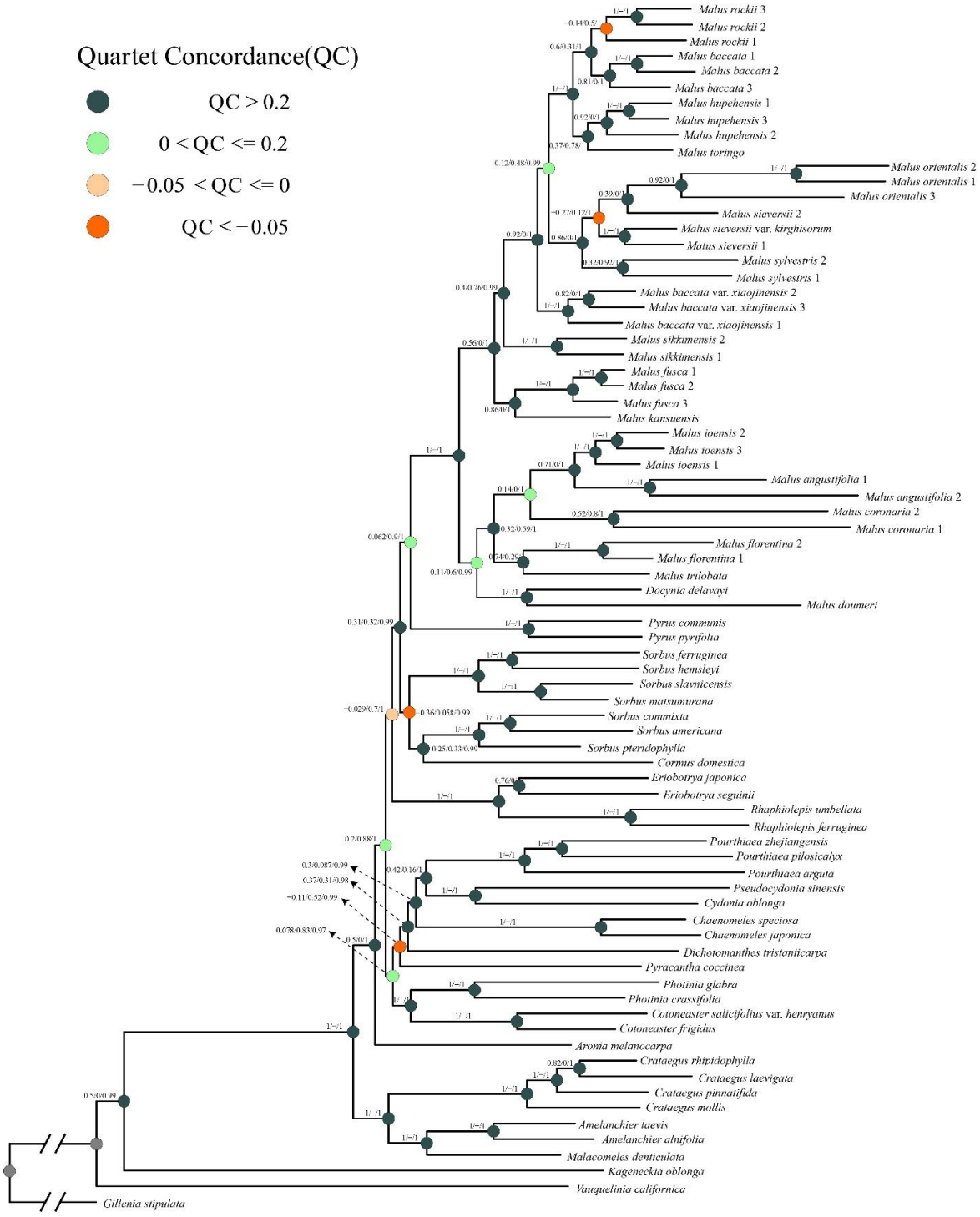

Figure S11. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80%-sample dataset

Quartet Samping Scores are shown on branches indicating Quartet Concordance/Quartet Differential/Quartet Informativeness. Quartet Concordance is color-coded according to the legend.

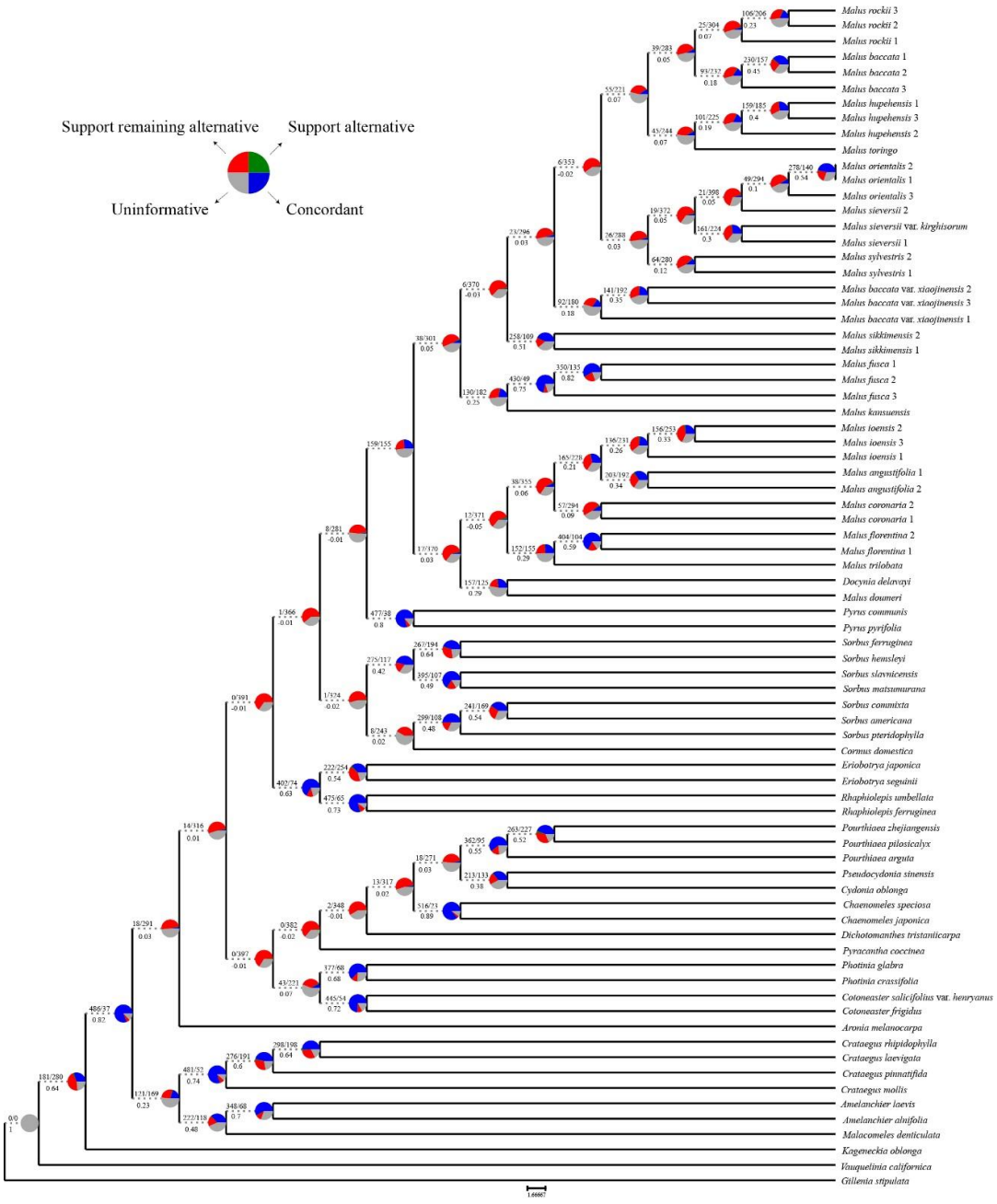

Figure S12. Maximum likelihood cladogram of *Malus* s.l. inferred from RAxML analysis of the concatenated 80%-sample dataset

Numbers above branches indicate the number of gene trees concordant/conflicting with that node in the species tree. Numbers below the branches are the Internode Certainty All (ICA) score. Pie charts on nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray).

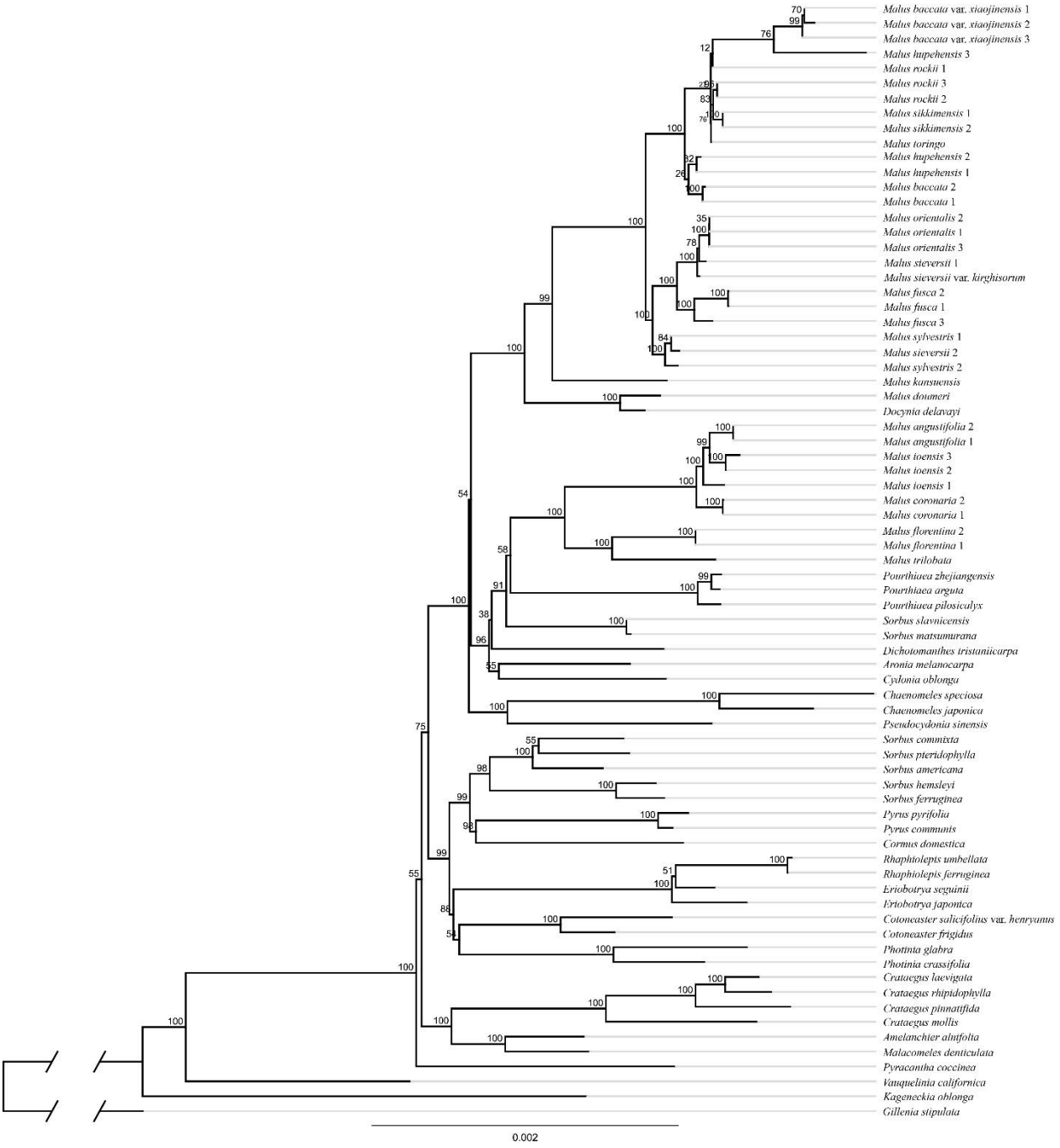

Figure S13. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80 plastid coding genes

Numbers above the branches indicate the bootstrap support.

228

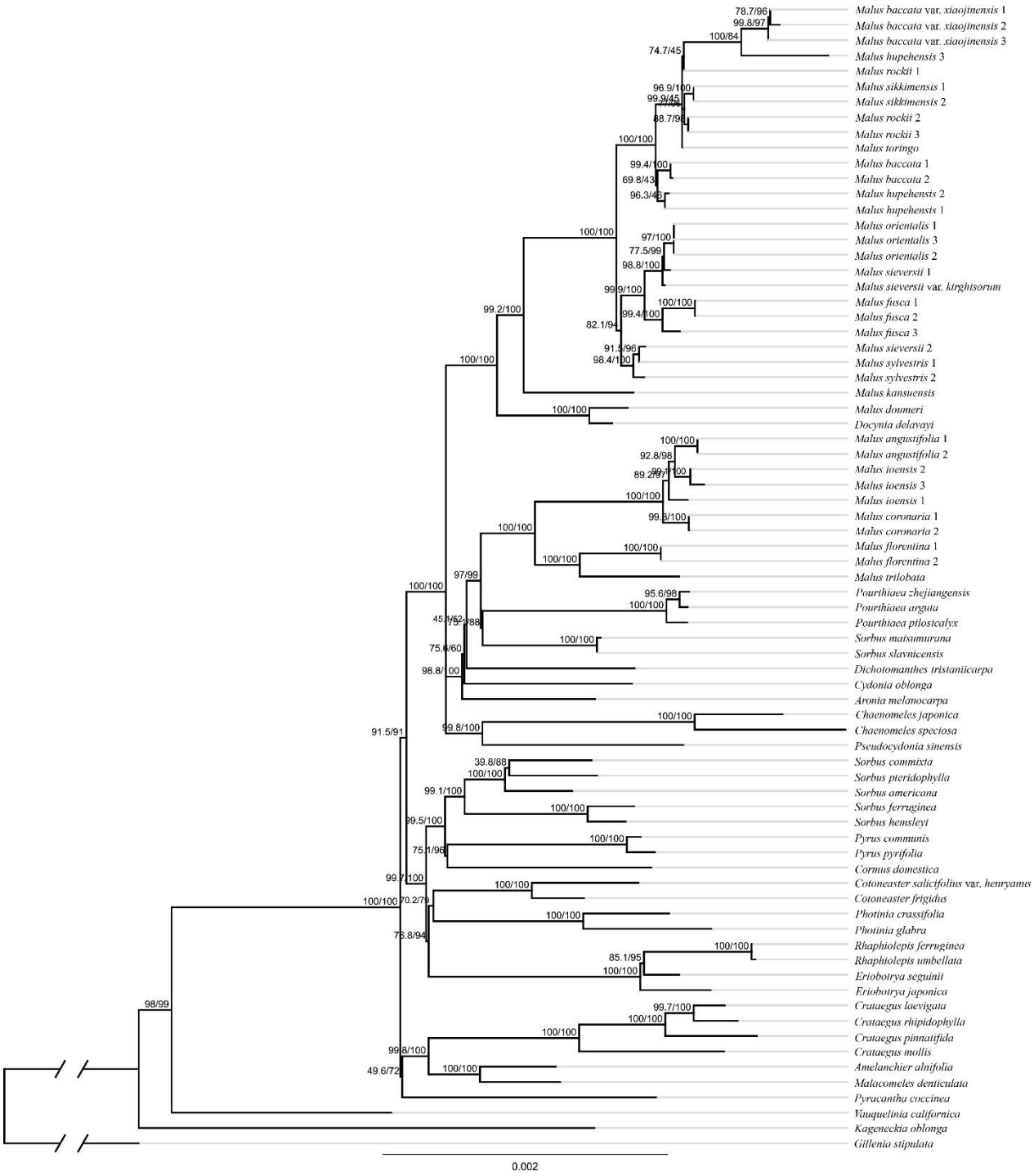

Figure S14. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from IQ-TREE2 analysis of the concatenated 80 plastid coding genes

Numbers above the branches indicate the SH-aLRT support and Ultrafast Bootstrap support.

229

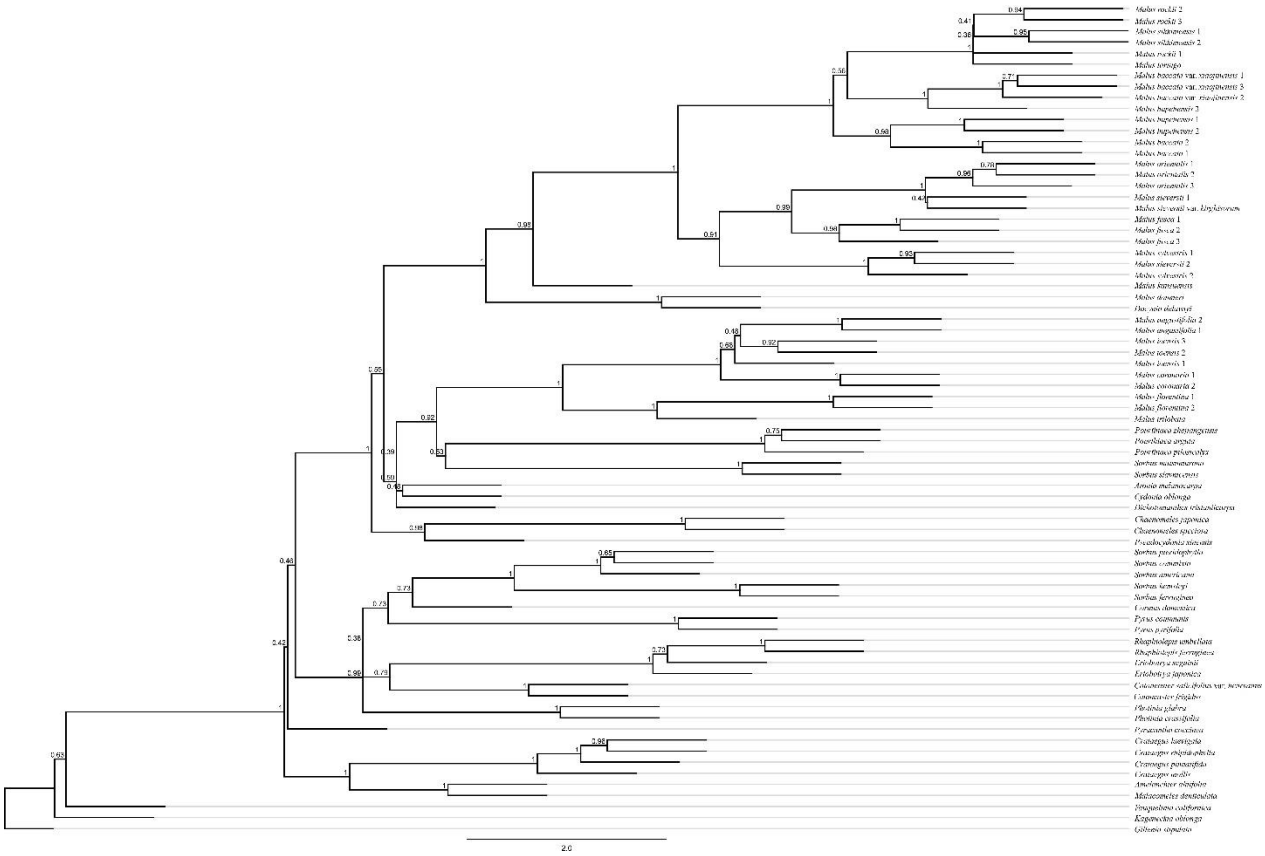

Figure S15. Species tree of *Malus* s.l. in the framework of Maleae inferred from ASTRAL-III of the concatenated 80 plastid coding genes

Numbers above the branches indicate the branch support values measuring the support for a local posterior possibility.

230

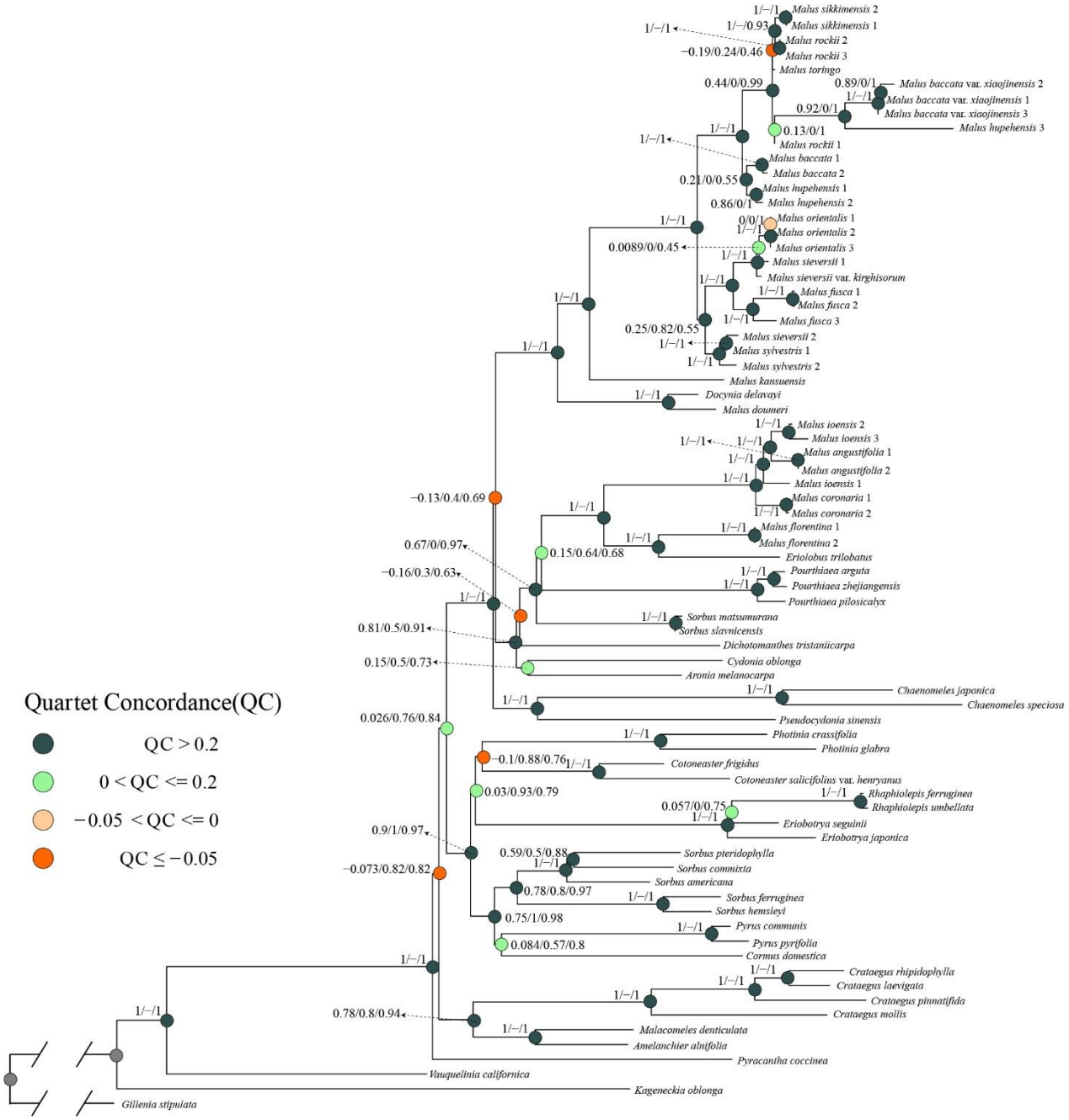

Figure S16. Maximum likelihood phylogeny of *Malus* s.l. in the framework of Maleae inferred from RAxML analysis of the concatenated 80 plastid coding genes

Quartet Samping Scores are shown on branches indicating Quartet Concordance/Quartet Differential/Quartet Informativeness. Quartet Concordance is color-coded according to the legend.

231

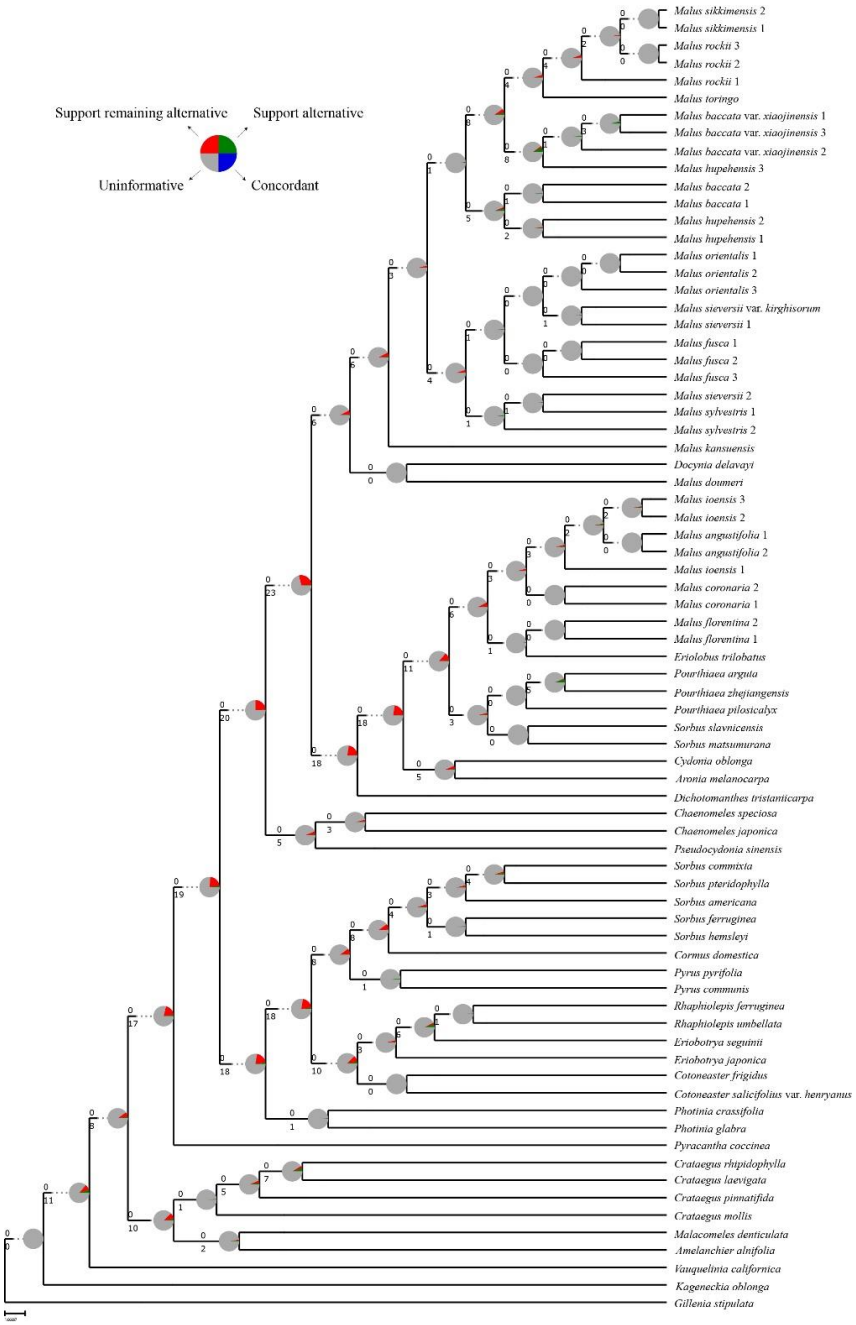

Figure S17. Maximum likelihood cladogram of *Malus* s.l. inferred from RAxML analysis of the concatenated 80 plastid coding genes

Numbers above branches indicate the number of gene trees concordant/conflicting with that node in the species tree. Numbers below the branches are the Internode Certainty All (ICA) score. Pie charts on nodes present the proportion of gene trees that support that clade (blue), the proportion that support the main alternative bifurcation (green), the proportion that support the remaining alternatives (red), and the proportion (conflict or support) that have < 50% bootstrap support (gray).

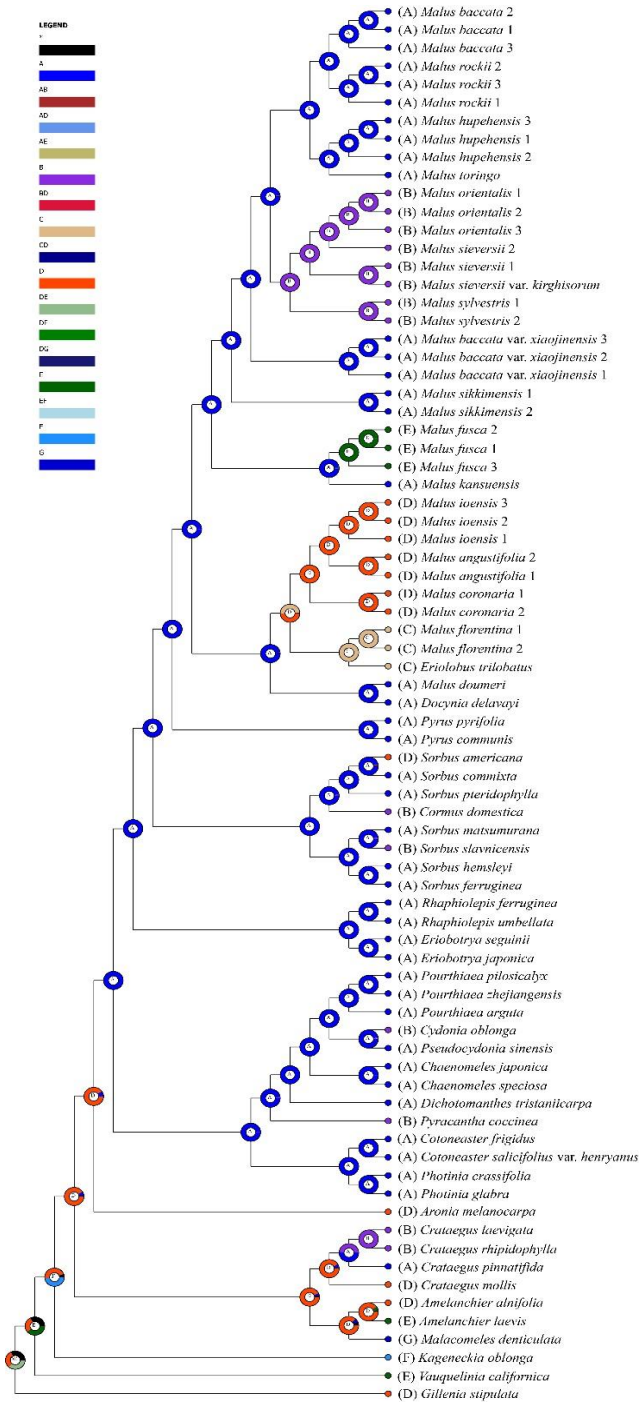

Figure S18. The ancestral area reconstruction using BioGeoBEARS from the nuclear SCN dataset with fossil record only used for calibration in the framework of Maleae, with the colored key identifying extant and possible ancestral ranges

(**A**), Southern East Asia; (**B**), Northern East Asia; (**C**), Europe excluding the Mediterranean; (**D**), Central Asia; (**E**), Mediterranean Europe/West Asia; (**F**), eastern North America; (**G**), western North America.

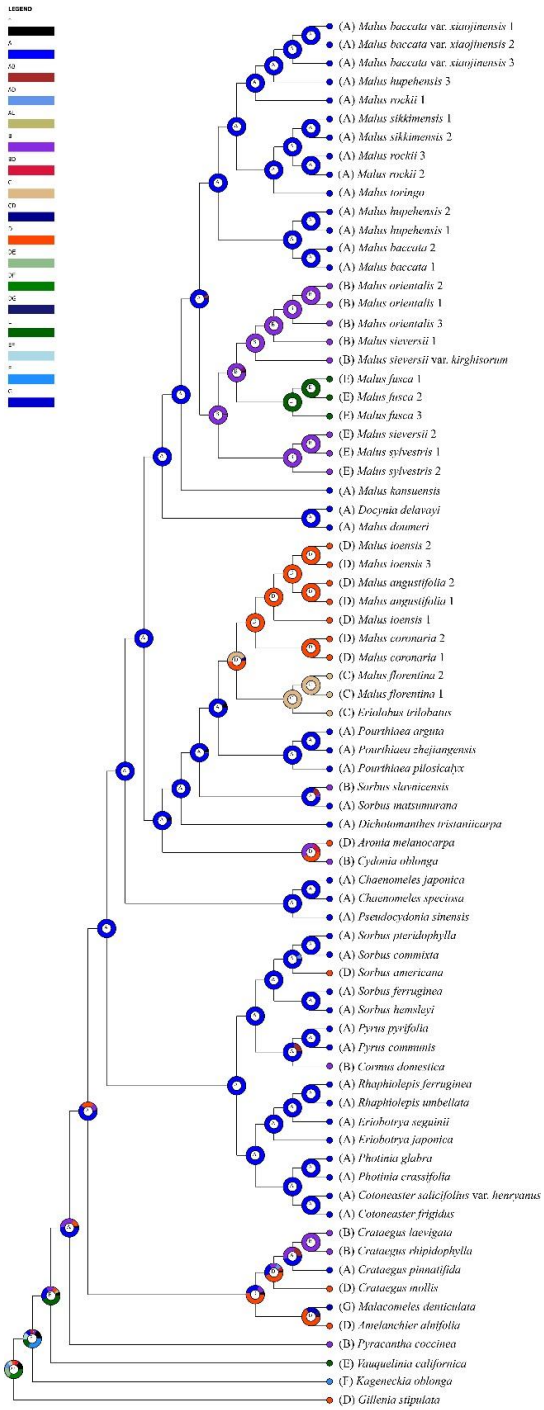

Figure S19. The ancestral area reconstruction using BioGeoBEARS from the 80 plastid coding genes dataset with fossil record only used for calibration in the framework of Maleae, with the colored key identifying extant and possible ancestral ranges
(**A**), Southern East Asia; (**B**), Northern East Asia; (**C**), Europe excluding the Mediterranean; (**D**), Central Asia; (**E**), Mediterranean Europe/West Asia; (**F**), eastern North America; (**G**), western North America.

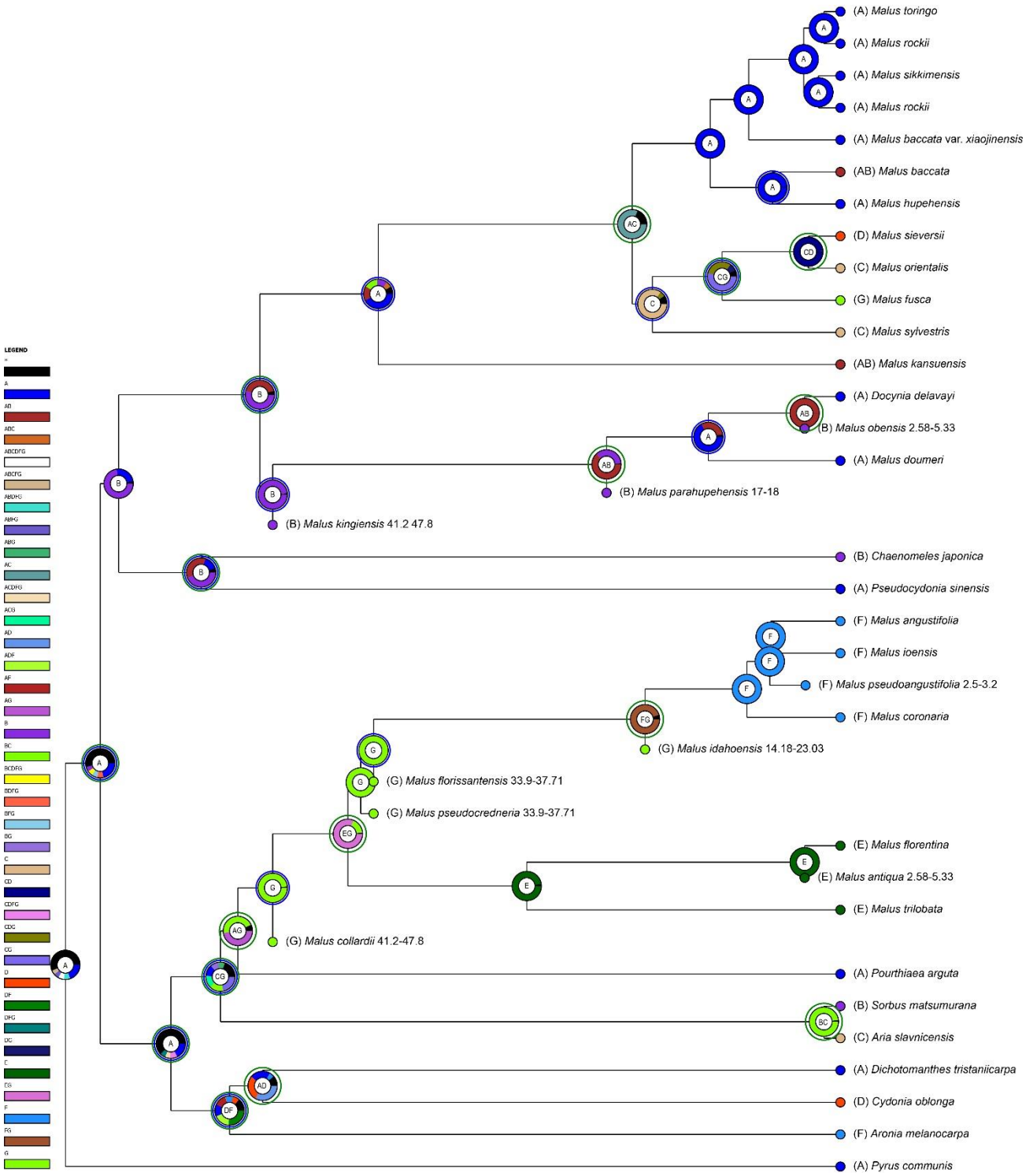

Figure S20. The ancestral area reconstruction using BioGeoBEARS from the 80 plastid coding genes dataset, with the colored key identifying extant and possible ancestral ranges (**A**), Southern East Asia; (**B**), Northern East Asia; (**C**), Europe; (**D**), Central Asia; (**E**), Mediterranean Europe/West Asia; (**F**), eastern North America; (**G**), western North America.

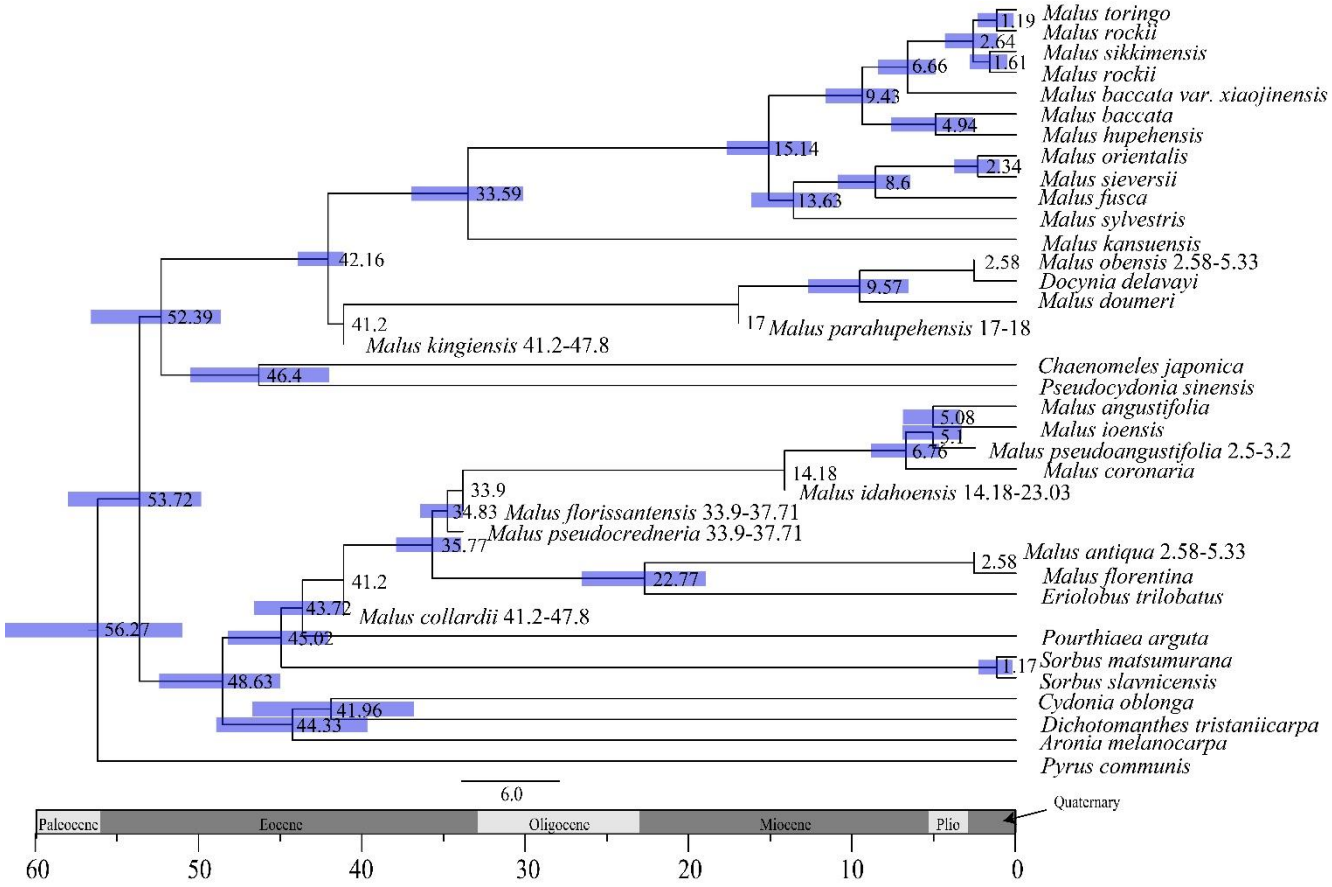

Figure S21. Dated chronogram for the apple genus *Malus* inferred from BEAST with the Fossilized Birth-Death process based on the 80 plastid codinig genes dataset

232

233 **Table S1.** Voucher and sequences information for taxa of the apple genus and its relatives used in this

234 study

235 **Table S2.** The number of pre-filtered and post-filtered sequences for each sample in this study

236 **Table S3.** The significant results from the HyDe analysis